ORIGINAL RESEARCH

Chapter 7: Grading a Body of Evidence on Diagnostic Tests

Sonal Singh, MD, MPH[1,2], Stephanie M. Chang, MD, MPH[3], David B. Matchar, MD[4,5], and Eric B. Bass, MD, MPH[1,2]

[1]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [2]Department of Epidemiology, Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD, USA; [3]Center for Outcomes and Evidence, Agency for Healthcare Research and Quality, Rockville, MD, USA; [4]Duke-NUS Medical School, Singapore, Singapore; [5]Duke Center for Clinical Health Policy Research, Durham, NC, USA.

INTRODUCTION: Grading the strength of a body of diagnostic test evidence involves challenges over and above those related to grading the evidence from health care intervention studies. This chapter identifies challenges and outlines principles for grading the body of evidence related to diagnostic test performance.

CHALLENGES: Diagnostic test evidence is challenging to grade because standard tools for grading evidence were designed for questions about treatment rather than diagnostic testing; and the clinical usefulness of a diagnostic test depends on multiple links in a chain of evidence connecting the performance of a test to changes in clinical outcomes.

PRINCIPLES: Reviewers grading the strength of a body of evidence on diagnostic tests should consider the principle domains of *risk of bias*, *directness*, *consistency, and precision*, as well as *publication bias*, *dose response association*, *plausible unmeasured confounders* that would decrease an effect, and *strength of association*, similar to what is done to grade evidence on treatment interventions. Given that most evidence regarding the clinical value of diagnostic tests is indirect, an analytic framework must be developed to clarify the key questions, and strength of evidence for each link in that framework should be graded separately. However if reviewers choose to combine domains into a single grade of evidence, they should explain their rationale for a particular summary grade and the relevant domains that were weighed in assigning the summary grade.

KEY WORDS: grades; diagnostic tests; publication bias; health care intervention.
J Gen Intern Med 27(Suppl 1):S47–55
DOI: 10.1007/s11606-012-2021-9

INTRODUCTION

"Grading" refers to the assessment of the strength of the body of evidence supporting a given statement or conclusion rather than to the quality of an individual study.[1] Grading can be valuable for providing information to decisionmakers, such as guideline panels, clinicians, caregivers, insurers and patients who wish to use an evidence synthesis to promote improved patient outcomes.[1,2] In particular, such grades allow decisionmakers to assess the degree to which any decision can be based on bodies of evidence that are of high, moderate, or only low strength of evidence. That is, decisionmakers can make a more defensible recommendation about the use of the given intervention or test than they might make without the strength of evidence grade.

The Evidence-based Practice Center (EPC) Program supported by the Agency for Healthcare Research and Quality (AHRQ) has published guidance on assessing the strength of a body of evidence when comparing medical interventions.[1,3] That guidance is based on the principles identified by the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group[4–6]with minor adaptations for EPCs. It is important to distinguish between the quality of a study and the strength of a body of evidence on diagnostic tests as assessed by the GRADE and EPC approaches. EPCs consider *"The extent to which all aspects of a study's design and conduct can be shown to protect against systematic bias, nonsystematic bias, and inferential error"* as the *quality* or *internal validity* or *risk of bias* of an individual study.[7] In contrast to the GRADE approach, the EPC approach prefers to use the term "strength of evidence" instead of "quality of evidence" to describe the grade of an evidence base for a given outcome because the latter term is often equated with the quality of individual studies without consideration of the other domains for grading a body of evidence. An assessment of the *strength* of the entire body of evidence includes an assessment of the quality of an individual study along with other domains. Although the GRADE approach can be used to make judgments about the strength of an evidence base and the strength of recommendations, this chapter considers using GRADE as a tool for assessing only the strength of an evidence base.

When assessing the strength of an evidence base, systematic reviewers should consider four principle domains—*risk of bias, consistency, directness,* and *preci-*

sion.[5] Additionally, reviewers may wish to consider *publication bias* as a fifth principle domain as recently suggested by the GRADE approach.[6] Additional domains to consider are *dose-response association*, *existence of plausible unmeasured confounders,* and *strength of association (i.e., magnitude of effect*). Of note, GRADE considers applicability as an element of *directness*. This is distinct from the EPC approach, which encourages users to evaluate applicability as a separate component.

EPCs grade the strength of evidence for each of the relevant outcomes and comparisons identified in the key questions addressed in a systematic review. The process of defining the important intermediate and clinical outcomes of interest for diagnostic tests is further described in a previous article.[8] Because most diagnostic test literature focuses on test performance (e.g., sensitivity and specificity), at least one key question will normally relate to that evidence. In the uncommon circumstance in which a diagnostic test is studied in the context of a clinical trial (e.g., test versus no test) with clinical outcomes as the study endpoint, the reader is referred to the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* on evaluating interventions.[1,3] For other key questions, such as those related to analytic validity, clinical validity, and clinical utility, the principles described in the present document and the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* should apply.

This paper is meant to complement the EPC Methods Guide for Comparative Effectiveness Reviews, and not to be a complete review. Although we have written this paper to serve as guidance for EPCs, we also intend for this to be a useful resource for other investigators interested in conducting systematic reviews on diagnostic tests. In this paper, we outline the particular challenges that systematic reviewers face in grading the strength of a body of evidence on diagnostic test performance. The focus of this article will be on diagnostic tests, meaning tests that are used in the diagnostic and management strategy of a patient symptom or complaint, as opposed to prognostic tests, which are for predicting responsiveness to treatment. We then propose principles for addressing these challenges.

COMMON CHALLENGES

Diagnostic test studies commonly focus on accuracy of the test to make a disease diagnosis, and the task of grading this body of evidence is a challenge in itself. Through discussion with EPC investigators and a review of recent EPC reports on diagnostic tests,[9–13] we identified common challenges that reviewers face when assessing the strength of a body of evidence on diagnostic test performance.

One common challenge is that standard tools for assessing the quality of a body of evidence associated with an intervention—in which the body of evidence typically relates directly to the overarching key question—are not so easily applied to a body of evidence associated with a diagnostic test, where evidence is often indirect. Indeed, this is the reason that establishing a logical chain with an analytic framework and the associated key questions is particularly important for evaluating a diagnostic test (*see* Paper 2).[8] It is also the reason we must assess the strength of the body of evidence for each link in the chain. The strength of the body of evidence regarding the overarching question of whether a test will improve clinical outcomes depends both on the total body of evidence, as well as the body of evidence for the weakest link in this chain. Although there is a temptation to use diagnostic accuracy as an intermediate outcome for the effect of a diagnostic test on clinical outcomes, there is often no direct linkage between the diagnostic accuracy outcome and a clinical outcome. This is particularly challenging when tests are used as a part of an algorithm. While rates of false positives and false negatives may be directly related to adverse effects or harms, other accuracy outcomes such as sensitivity or specificity may not directly correlate to effective management and treatment of disease, especially when the test under question is not directly linked to the use of an established treatment algorithm. When tests are used in regular practice not for final diagnosis and treatment, but as a triage for further testing, then accuracy of diagnosis is less important than accuracy of risk classification.

A second challenge arises in the application of the strength of evidence domains for studies of diagnostic tests. For example, in assessing the precision of estimates of test performance, it is particularly difficult to judge whether a particular confidence interval is sufficiently precise; because of the logarithmic nature of diagnostic performance measurements—such as sensitivity, specificity, likelihood ratios, and diagnostic odds ratios—even a relatively wide confidence interval suggesting imprecision may not necessarily translate into imprecision that is clinically meaningful. Table 1 shows an example where a 10% reduction in the sensitivity of various biopsy techniques (from 98% to 88% in the far right column) changes the estimated probability of having cancer after a negative test by less than 5%.[11]

PRINCIPLES FOR ADDRESSING THE CHALLENGES

Principle 1. Methods for grading intervention studies can be adapted for studies evaluating studies on diagnostic tests with clinical outcomes

A body of evidence evaluating diagnostic test outcomes such as diagnostic thinking, therapeutic choice, and clinical

Table 1. Example of the Impact of Precision of Sensitivity on Negative Predictive Value

Type of biopsy	Post biopsy probability of having cancer after a negative core-needle biopsy result[a]			
	Analysis results	Analysis overestimated sensitivity by 1% (e.g., sensitivity 97% rather than 98%)	Analysis overestimated sensitivity by 5% (e.g., sensitivity 93% rather than 98%)	Analysis overestimated sensitivity by 10% (e.g., sensitivity 88% rather than 98%)
Freehand automated gun	6%	6%	8%	9%
Ultrasound guidance automated gun	1%	1%	3%	5%
Stereotactic guidance automated gun	1%	1%	3%	5%
Ultrasound guidance vacuum-assisted	2%	2%	3%	6%
Stereotactic guidance vacuum-assisted	0.4%	0.8%	3%	5%

[a]*For a woman with a BI-RADS® 4 score following mammography and expected to have an approximate prebiopsy risk of malignancy of 30%. Note that an individual woman's risk may be different from these estimates depending on her own individual characteristics*[71]

outcomes can be assessed in very much the same way as a body of evidence evaluating outcomes of therapeutic interventions. Grading issues in this type of diagnostic test study are more straightforward than in studies measuring accuracy outcomes. Although this is rarely done, the effect of tests on the clinical outcomes described above can be assessed directly with trial evidence. In cases where trial evidence is available, application of the grading criteria, such as GRADE, should not significantly differ from the methods used for intervention evidence.

An unresolved issue is what to do when there is no direct evidence available linking the test to the outcome of interest. For grading intervention studies, the use of intermediate outcomes, such as accuracy outcomes, would be considered "indirect" evidence and would reduce the strength of the grade. The linkage of accuracy outcomes such as true positives and false positives to clinical outcomes depend in part upon the benefits and harms of available treatments as well as the cognitive or emotional outcomes resulting from the knowledge itself, as outlined in a previous article [ref Segal et al.].[14]

Currently there is no consensus for one particular approach to grading an overall body of evidence when it is entirely indirect, such as when only studies of accuracy are available. As discussed in a previous article,[8] there are circumstances in which accuracy outcomes may be sufficient to conclude that there is or is not a benefit on clinical outcomes.[15] In other cases in which only indirect evidence on intermediate accuracy outcomes is available, EPCs should discuss with decisionmakers and methodologists the benefits of including such indirect evidence and the specific methods to be used.

Principle 2. Consider carefully what test characteristic measures are the most appropriate intermediate outcomes for assessing the impact of a test on clinical outcomes and their precision in the clinical context represented by the key question.

Consistent with EPC and GRADE principles of emphasizing the patient important outcomes, reviewers should consider how any surrogates such as accuracy outcomes will lead to changes in clinical outcomes. Use of an analytic framework and decision models as described in paper two [ref Samson et al.],[8] help to clarify the linkage between accuracy outcomes and clinical outcomes for systematic reviewers, and users of systematic reviews alike.

If accuracy outcomes are presented as true positives, true negatives, false positives, and false negatives, then they can be easily translated into other accuracy outcomes such as sensitivity and specificity, positive predictive value (PPV) and negative predictive value (NPV). Systematic reviewers need to carefully consider which of these accuracy outcomes to assess based on which outcome will relate most directly to clinical outcomes as well as necessary levels of precision.

Sometimes it is more important to "rule out" a particular disease that has severe consequences if missed. In these cases, use of a triage test with high sensitivity and NPV may be what is most important, and actual diagnosis of a particular disease is less important.

When the treatment of a disease has high associated risks, multiple tests are often used to assure the highest accuracy. Tests used in isolation need to have both high sensitivity and specificity, or high PPV and NPV, but if no such test is available, clinicians may be interested in the added benefits and harms of "adding-on" a test. The accuracy outcome of interest of these tests would primarily be high specificity or PPV.

Tests that are more invasive will naturally have greater harms. Additional harms may result from misdiagnosis, so it is almost always important to consider the measurement of false positives and false negatives when assessing the harms of a diagnostic test. The degree of harms from false negatives depend on the severity of disease if there is a missed diagnosis, in addition to the risks from the testing itself (i.e. if the test is invasive and associated with risks in and of itself). The degree of harms from false positives depends on the invasiveness of further testing or treatment, as well as the emotional and cognitive effects of inaccurate disease labeling.

As a simple example, one might have compelling data regarding the value of outcomes resulting from true positive test result, as well as true negative, false positive and false negative results. In a simple decision model it is possible to identify a threshold line for the combinations of test sensitivity and specificity for which testing vs. not testing is a toss-up—where net benefits are equivalent to net harms. To the extent that the confidence intervals for sensitivity and specificity derived from the body of evidence are contained within one territory or the other ("testing better", as in this illustration), these intervals are sufficiently precise for purposes of decision making.[16]

Of course, this formulation is a simplification for many situations. Tests are rarely used alone to diagnosis disease and determining treatment choices, but are more often used as part of an algorithm of testing and management. The accuracy outcome of most interest depends on how the test is used in a clinical algorithm, as well as the mechanism by which the test could improve clinical outcomes or cause harms. Whether or not one uses a decision model to help sort out these issues, considering the important test characteristics and their precision in the clinical context represented by the key question is a necessary step in the process of assessing a body of evidence.

Principle 3. The principle domains of GRADE can be adapted to assess a body of evidence on diagnostic test accuracy.

To assess a body of evidence related to diagnostic test performance, we can adapt the GRADE's principle domains of *risk of bias, consistency, directness,* and *precision*. (Table 2) Evaluating *risk of bias* includes considerations of how the study type and study design and conduct may have contributed to systematic bias. The potential sources of bias relevant to diagnostic test performance and strategies for assessing the risk of systematic error in such studies are discussed in a previous article [ref Santaguida et al.].[17] Diagnostic tests, particularly laboratory tests, can yield heterogeneous results due to different technical methods. For example, studies may report using different antibodies for immunoassays, or standards with different values and units assigned to them.

Consistency concerns homogeneity in the direction and magnitude of results across different studies. The concept can be similarly applied to diagnostic test performance studies, although the method of measurement may differ. For example, consistency among intervention studies with quantitative data may be assessed visually with a forest plot. However, for diagnostic test performance reviews, the most common presentation format is a summary receiver operating characteristic (ROC) curve, which displays the sensitivity and specificity results from various studies. A bubble plot of true positive versus false positive rates showing spread in ROC space is one method of assessing the consistency of diagnostic accuracy among studies. As with intervention studies, the strength of evidence is reduced by unexplained heterogeneity—that is, heterogeneity not explained by different study designs, methodologic quality of studies, diversity in subject characteristics, or study context.

Directness, according to AHRQ's *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*,[3] occurs when the evidence being assessed "reflects a single, direct link between the interventions of interest [diagnostic tests] and the ultimate health outcome under consideration."[1] When assessing the directness of the overarching question, if there are no studies linking the test to a clinical outcome, then evidence that only provides diagnostic accuracy outcomes would be considered indirect. If the decision is made to grade the strength of evidence of an intermediate outcome such as diagnostic accuracy, then the reviewer does not need to automatically "downgrade" that outcome for being indirect. Of note, directness does apply to how a test is used in comparison to another test. For example, a study may compare the use of a d-dimer test as a replacement to venous ultrasound for the diagnosis of venous thromboembolism, but in actual practice the relevant question may be the comparison of d-dimer test as a triage for venous ultrasound compared to the use of ultrasound alone. It is worth noting EPCs consider some aspects of directness separately as described in the applicability chapter [Hartmann et al.].[18] Although not included when EPCs assess directness or the strength of evidence, other schemas, such as GRADE, rate directness based on whether the test evaluated is not the exact test used in practice, or if the test accuracy is being calculated in a population or for a use (diagnosis, prognosis, etc.) that is different than the population or use evaluated in the report. Because EPC reports are intended to be used by a broad spectrum of stakeholders, describing the applicability of the evidence on these factors allows the decision-makers to consider how the evidence relates to their test and population.

Precision refers to the width of confidence intervals for diagnostic accuracy measurements and is integrally related to sample size.[1] Before downgrading the strength of an evidence base for imprecision, reviewers could consider how imprecision for one measure of accuracy may impact clinically meaningful outcomes. This may involve a simple calculation of posttest probabilities over a range of values for sensitivity and specificity, as shown in Table 1, or, as illustrated above, a more formal analysis with a decision model (*see* Trikalinos et al.).[19] If the impact of imprecision on clinical outcomes is negligible or if the demonstrated precision is sufficient to make the decision, the evidence should not be downgraded.

Table 2. Required and Additional Domains and their Definitions*

Domain	Definition and Elements	Application to evaluation of diagnostic test performance
Risk of Bias	Risk of bias is the degree to which the included studies for a given outcome or comparison have a high likelihood of adequate protection against bias (i.e., good internal validity), assessed through main elements: • Study design (e.g., RCTs or observational studies) • Aggregate quality of the studies under consideration from the rating of quality (good/fair/poor) done for individual studies	Use one of three levels of aggregate risk of bias: • Low risk of bias • Medium risk of bias • High risk of bias Well designed and executed studies of new tests compared to an adequate criterion standard are rated as "low risk of bias"
Consistency	Consistency is the degree to which reported study results (e.g., sensitivity, specificity, likelihood ratios) from included studies are similar. This can be assessed through two main elements: • The range of study results is narrow • *Variability in study results is explained by differences in study design, patient population or test variability*	Use one of three levels of consistency: • Consistent (i.e., no inconsistency) • Inconsistent • Unknown or not applicable (e.g., single study) Single-study evidence bases should be considered as "consistency unknown (single study)."
Directness	Directness relates to whether the evidence links the interventions directly to outcomes. For a comparison of two diagnostic tests, directness implies head-to-head comparisons against a common criterion standard Directness may be contingent on the outcomes of interest	Score dichotomously as one of two levels of directness • Direct • Indirect When assessing the directness of the overarching question, if there are no studies linking the test to a clinical outcome, then evidence that only provides diagnostic accuracy outcomes would be considered indirect. If indirect, specify which of the two types of indirectness account for the rating (or both, if that is the case)—namely, use of intermediate/ surrogate outcomes rather than health outcomes, and use of indirect comparisons. If the decision is made to grade the strength of evidence of an intermediate outcome such as diagnostic accuracy, then the reviewer does not need to automatically "downgrade" that outcome for being indirect
Precision	Precision is the degree of certainty surrounding an effect estimate with respect to a given outcome (i.e., for each outcome separately) If a meta-analysis was performed, this will be the confidence interval around the summary measure(s) of test performance (e.g sensitivity, true positive)	Score dichotomously as one of two levels of precision: • Precise • Imprecise A precise estimate is an estimate that would allow a clinically useful conclusion. An imprecise estimate is one for which the confidence interval is wide enough to include clinically distinct conclusions
Publication bias†	Publication bias indicates that studies may have been published selectively, with the result that the estimate of test performance based on published studies does not reflect the true effect. Methods to detect publication bias for medical test studies are not robust. Evidence from small studies of new tests or asymmetry in funnel plots should raise suspicion for publication bias	Publication bias can influence ratings of consistency, precision, magnitude of effect (and, to a lesser degree, risk of bias and directness). Reviewers should comment on publication bias when circumstances suggest that relevant empirical findings, particularly negative or no-difference findings, have not been published or are unavailable
Dose-response association	This association, either across or within studies, refers to a pattern of a larger effect with greater exposure (dose, duration, and adherence)	The dose-response association may support an underlying mechanism of detection and potential relevance for some tests that have continuous outcomes and possibly multiple cutoffs [e.g., gene expression, serum PSA (prostate-specific antigen) levels, and ventilation/perfusion scanning]
Plausible unmeasured confounding and bias that would decrease an observed effect or increase an effect if none was observed	Occasionally, in an observational study, plausible confounding factors would work in the direction opposite to that of the observed effect. Had these confounders not been present, the observed effect would have larger. In such case the evidence can be upgraded	The impact of plausible unmeasured confounders may be relevant to testing strategies that predict outcomes. A study may be biased to find low diagnostic accuracy via spectrum bias and yet despite this find very high diagnostic accuracy

Table 2. (continued)

Domain	Definition and Elements	Application to evaluation of diagnostic test performance
Strength of association (magnitude of effect)	Strength of association refers to the likelihood that the observed effect or association is large enough that it cannot have occurred solely as a result of bias from potential confounding factors	The strength of association may be relevant when comparing the accuracy of two different medical tests with one being more accurate than the other It is possible that the accuracy of a test is better than the reference standard because of an imperfect reference standard. It is important to consider this and modify the analysis to take into consideration alternative assumptions about the best reference standard

*Adapted from the Methods Guide for Effectiveness and Comparative Effectiveness Reviews[3]
†The GRADE approach is moving towards considering publication bias a GRADE principle domain
Abbreviations: EPC = Evidence-based Practice Center

Principle 4. Additional GRADE domains can be adapted to assess a body of evidence on diagnostic test accuracy.

When grading a body of evidence about a diagnostic test, additional domains should be considered. These additional domains are summarized in Table 2.[1–3] These additional domains include *publication bias, dose-response association, existence of plausible unmeasured confounders,* and *strength of association.* Reviewers should comment on *publication bias* when circumstances suggest that negative or no-difference findings have not been published or are unavailable. The *dose-response association* may support an underlying mechanism of detection and potential relevance for some tests that have continuous outcomes and possibly multiple cutoffs (e.g., gene expression, serum PSA [prostate-specific antigen] levels, and ventilation/perfusion scanning). The impact of *plausible unmeasured confounders* may be relevant to testing strategies that predict outcomes. A study may be biased to find low diagnostic accuracy via spectrum bias and yet despite this find very high diagnostic accuracy. The *strength of association* may be relevant when comparing the accuracy of two different diagnostic tests with one being more accurate than the other.

Principle 5. Multiple domains should be incorporated into an overall assessment in a transparent way

The overall strength of evidence reflects a global assessment of the principle domains and any additional domains, as needed, into an overall summary grade—high, moderate, low, or insufficient evidence. The focus should be on providing an overall grade for the relevant key question link in the analytic chain or for outcomes considered relevant for patients and decisionmakers. These should ideally be identified a priori. Consideration should be given on how to incorporate multiple domains into the overall assessment.

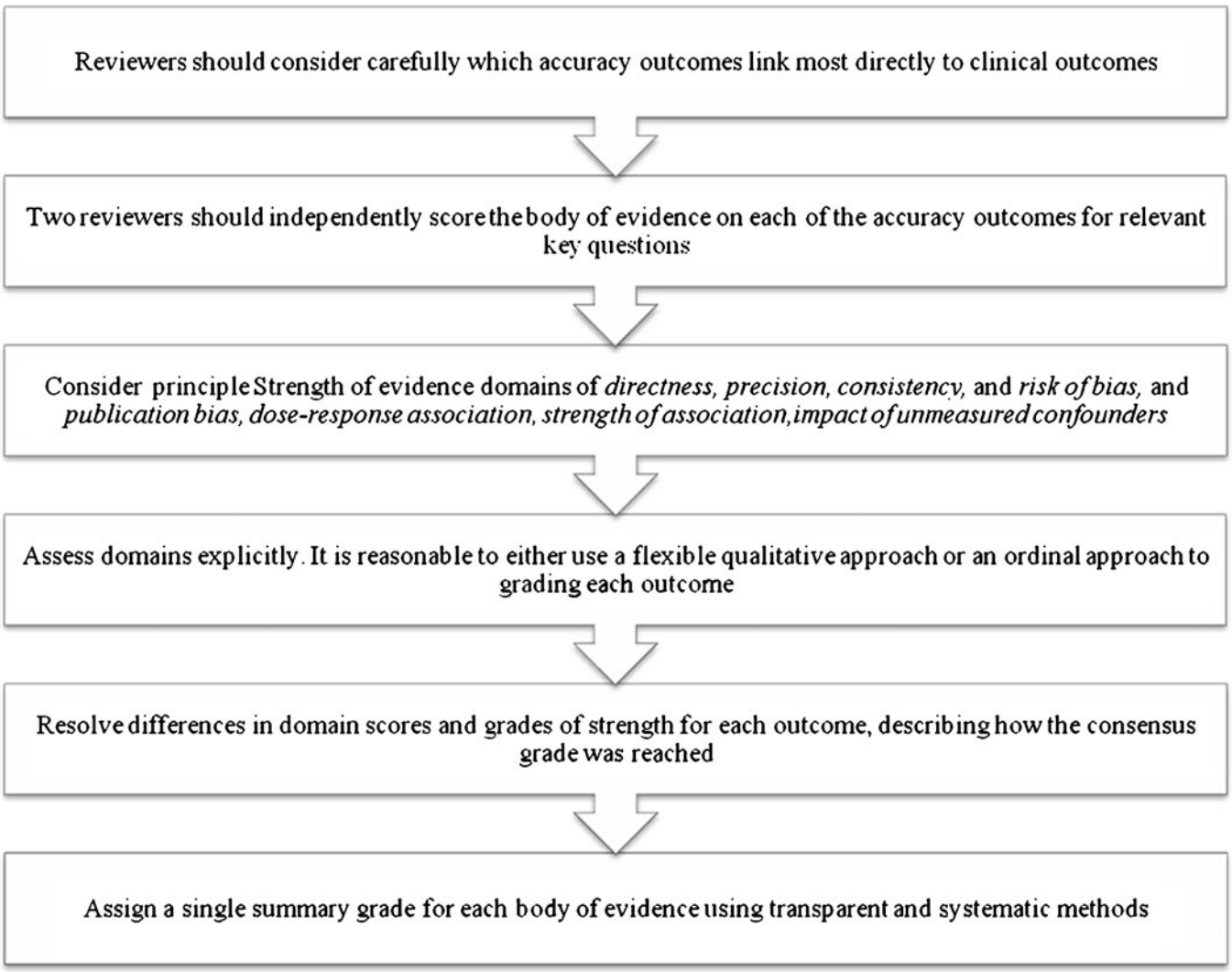
There is no empirical evidence to suggest any difference in assigning a summary grade based on qualitative versus quantitative approaches. GRADE advocates an ordinal approach with a ranking from high, moderate, or low, to very low. These "grades" or "overall ratings" are developed using the eight domains suggested by GRADE. The EPC approach for intervention studies described in the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*[1,3] allows for more flexibility on grading the strength of evidence. Whichever approach reviewers choose for diagnostic tests, they should consider describing their rationale for which of the required domains were weighted the most in assigning the summary grades.

ILLUSTRATION

An illustration in Table 3 provides guidance on how reviewers should approach grading a body of evidence on diagnostic test accuracy. This is adapted from the GRADE approach and the EPC Methods Guide for Comparative Effectiveness Reviews. Reviewers should carefully consider which accuracy outcomes are linked to clinical outcomes. In choosing the accuracy outcomes, if the diagnostic test is followed by an invasive procedure, then the number of false positives may be considered most important. However when "diagnostic tests" are used as part of a management strategy, consideration should also be given to grading the positive predictive value and the negative predictive value or likelihood ratios if these additional outcomes assist decision makers. An additional example of grading for positive predictive value and negative predictive value is shown in the norovirus table below (Table 4).[20,21]This table illustrates how presentation of the same information in different ways can be helpful in considering how to link the accuracy outcomes to clinical practice and projecting how the test would impact clinical outcomes.

Another review on the use of non-invasive imaging in addition to standard workup after recall for evaluation of a breast lesion detected on screening mammography or physical examination illustrates how accuracy does not relate to outcomes when it is being used as part of an algorithm of whether to treat versus watchful waiting.[13] This evidence review focused on the non-invasive imaging studies intended to guide patient management decisions after the discovery of a possible abnormality. The studies were intended to provide additional information to enable

Table 3. Steps in Grading a Body of Evidence on Diagnostic Test Accuracy Outcomes



**Adapted from the Methods Guide for Effectiveness and Comparative Effectiveness Reviews*[3]

women to be appropriately triaged into "biopsy," "watchful waiting," or "return to normal screening intervals" care pathways. Thus the usual strategy of assuming the clinical outcome would be simply a downgrade of the surrogate doesn't always hold true. Reviewers should evaluate the surrogate in the context of the clinical outcome. As the table on summary of key findings in the evidence report illustrates, despite the accuracy of the exact diagnosis being

Table 4. Illustration of the Approach to Grading a Body of Evidence on Diagnostic Tests- Identifying Norovirus in a Healthcare Setting *

Outcome	Quantity and type of evidence	Findings	Starting grade	Decrease GRADE‡					GRADE of Evidence for Outcome	Overall GRADE§
				Risk of Bias‡	Consistency‡	Directness‡	Precision‡	Publication Bias‡		
Sensitivity†	1 DIAG	68%	High	0	0	0	-1	0	Moderate	Moderate
Specificity†	1 DIAG	99%	High	0	0	0	-1	0	Moderate	
PPV†	1 DIAG	97%	High	0	0	0	-1	0	Moderate	
NPV†	1 DIAG	82%	High	0	0	0	-1	0	Moderate	

**Adapted from MacCannell T, Umscheid CA, Agarwal RK, Lee I, Kuntz G, Stevenson, KB, and the Healthcare Infection Control Practices Advisory Committee. Guideline for the prevention and control of norovirus gastroenteritis outbreaks in healthcare settings. Infection Control and Hospital Epidemiology. 2011; 32([10]): 939-969*[20][21]

†These outcomes were considered the most critical by the guideline developers
‡These modifiers can impact the GRADE by 1 or 2 points
§Consider the additional domains of strength of association, dose-response and impact of plausible confounders if applicable

low, clinical management may be the same if the post-test probability did not cross a certain decision threshold to alter management decisions.

Two reviewers should independently score these relevant major outcomes and comparisons, within each key question. They should consider the principle domains of directness, precision, consistency, risk of bias, and publication bias, as well as dose response association, strength of association, and impact of unmeasured confounders. Reviewers should explicitly assess each domain to arrive at a grade for each outcome. Reviewer's choice of various accuracy outcomes to grade may affect how the various domains of *directness, precision* and *consistency* are assessed. This is illustrated in the example by the GRADE working group about multi-splice coronary CT scans as compared to coronary angiography.[4] Evidence was considered direct for certain accuracy outcome such as true positives, true negatives, and false positives since there was little uncertainity about the clinical implications of these results. However, since there was uncertainity about the clinical implications of a false negative test result, this was considered indirect.[4] This resulted in a low strength of evidence grade for false negatives as compared to moderate for other accuracy outcomes.

It is reasonable to consider either a more flexible qualitative approach to grading or the standard ordinal approach ranging from high to insufficient strength of evidence. Reviewers should resolve differences in domain assessments and grades of outcomes and describe how the consensus score was reached (e.g., by discussion or by third-party adjudication). If appropriate they should consider arriving at a single summary grade for the diagnostic test through transparent and systematic methods. If reviewers chose to assign an overall summary grade they should consider the impact of various accuracy outcomes on the overall strength of evidence grade and identify which of these accuracy outcomes was considered "key".

SUMMARY

Grading the strength of a body of diagnostic test evidence involves challenges over and above those related to grading the evidence from therapeutic intervention studies. The greatest challenge appears to be assessing multiple links in a chain of evidence connecting the performance of a test to changes in clinical outcomes. In this chapter, we focused primarily on grading the body of evidence related to a crucial link in the chain—diagnostic test performance—and described less fully the challenges involved in assessing other links in the chain.

No one system for grading the strength of evidence for diagnostic tests has been shown to be superior to any other and many are still early in development. However, we conclude that, in the interim, applying the consistent and transparent system of grading using the domains described above, and giving an explicit rationale for the choice of grades based on these domains, will make EPC and other reports on diagnostic tests more useful for decisionmakers.

KEY POINTS

- One can use GRADE for diagnostic tests. The outcomes one should consider are the clinical outcomes of effectiveness or harm if available for diagnostic tests.
- When intermediate accuracy outcomes are used, an analytic framework should describe how the test is related to clinical outcomes, and then delineate the individual questions that can be answered in that framework
- Selection of accuracy outcomes (i.e. sensitivity, specificity, positive predictive value and negative predictive value, true positives, true negatives, false positives, and false negatives) and needed levels of precision of these quantities should consider how the test is to be used in the clinical context.
- Domains of risk of bias, directness, consistency, precision, publication bias, dose response association, and plausible unmeasured confounders can be used to grade the strength of evidence for the effect of a diagnostic test on clinical outcomes or on intermediate surrogate outcomes if selected by the EPC and key informants.
- Whether reviewers choose a qualitative or quantitative approach to combining domains into a single grade, they should consider explaining their rationale for a particular summary grade and the relevant domains that were weighted the most in assigning the summary grade.

ACKNOWLEDGMENTS: *This report is based on research conducted by the Johns Hopkins Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD. The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services. The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment. This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products or actions may not be stated or implied.*

The authors would like to acknowledge the contribution of Dr Mark Helfand (Oregon Evidence-based Practice Center), Dr Joseph Lau (Tufts Evidence-based Practice Center) Dr Jonathan Treadwell (ECRI Institute Evidence-based Practice Center), Dr Kathleen N. Lohr (RTI International) and Dr Douglas K Owens (Stanford University) for providing comments on a draft of the manuscript.

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

Corresponding Author: *Sonal Singh, MD, MPH; Department of Medicine, Johns Hopkins University School of Medicine, 624 N Broadway, Rm 680 B, Baltimore, MD 21205, USA (e-mail: Ssingh31@jhu.edu).*

REFERENCES

1. **Owens DK, Lohr KN, Atkins D, et al.** AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions–Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol. 2010;63(5):513–23.
2. **Atkins D, Fink K, Slutsky J.** Better information for better health care: the evidence-based practice center program and the Agency for Healthcare Research and Quality. Ann Intern Med. 2005;142(12 Pt 2):1035–41.
3. Agency for Healthcare Research and Quality. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville, MD: Agency for Healthcare Research and Quality. Available at: http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318. December, 2011.
4. **Schunemann HJ, Oxman AD, Brozek J, et al.** Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336(7653):1106–10.
5. **Guyatt GH, Oxman AD, Vist GE, et al.** GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6.
6. **Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH.** GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401–6.
7. **Lohr KN, Carey TS.** Assessing "best evidence": issues in grading the quality of studies for systematic reviews. Jt Comm J Qual Improv. 1999;25:470–9.
8. **Samson D, Schoelles KM.** Chapter 2: Medical tests guidance (2) developing the topi and structuring systematic reviews of medical tests: utility of PICOTS, analytic frameworks, decision trees, and other frameworks. J Gen Internal Med. 2012. doi:10.1007/s11606-012-2007-7.
9. **Marchionni L, Wilson RF, Marinopoulos SS, et al.** Impact of gene expression profiling tests on breast cancer outcomes. Evidence report/technology assessment No. 160. (Prepared by The Johns Hopkins University Evidence-based Practice Center under contract No. 290-02-0018). AHRQ Publication No. 08-E002. Rockville, MD: Agency for Healthcare Research and Quality. January 2008. Available at: www.ahrq.gov/downloads/pub/evidence/pdf/brcancergene/brcangene.pdf. Accessed December, 2011.
10. **Ross SD, Allen IE, Harrison KJ, et al.** Systematic review of the literature regarding the diagnosis of sleep apnea. Evidence report/technology assessment No. 1. (Prepared by MetaWorks Inc. under Contract No. 290-97-0016.) AHCPR Publication No. 99-E002. Rockville, MD: Agency for Health Care Policy and Research. February 1999. Available at: www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=erta1. Accessed December, 2011.
11. **Bruening W, Schoelles K, Treadwell J, et al.** Comparative effectiveness of core-needle and open surgical biopsy for the diagnosis of breast lesions. Comparative effectiveness review No. 19. (Prepared by ECRI Institute Evidence-based Practice Center under Contract No. 290-02-0019.) Rockville, MD: Agency for Healthcare Research and Quality. December 2009. Available at: http://effectivehealthcare.ahrq.gov/ehc/products/17/370/finalbodyforposting.pdf. Accessed December, 2011.
12. **Segal JB, Brotman DJ, Emadi A, et al.** Outcomes of genetic testing in adults with a history of venous thromboembolism. Evidence report/technology assessment No. 180. (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. HHSA 290-2007-10061-I). AHRQ Publication No. 09-E011. Rockville, MD. Agency for Healthcare Research and Quality. June 2009. Available at: http://www.ahrq.gov/downloads/pub/evidence/pdf/factorvleiden/fvl.pdf. Accessed December, 2011.
13. **Bruening W, Uhl S, Fontanarosa J, Reston J, Treadwell J, Schoelles K.** Noninvasive Diagnostic Tests for Breast Abnormalities: Update of a 2006 Review. Comparative Effectiveness Review No. 47. (Prepared by the ECRI Institute Evidence-based Practice Center under Contract No. 290-02-0019.) AHRQ Publication No. 12-EHC014-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2012.
14. **Segal JB.** Chapter 3: Choosing the important outcomes for a systematic review of a medical test. J Gen Internal Med. 2011. doi:10.1007/s11606-011-1802-x.
15. **Lord SJ, Irwig L, Simes J.** When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need a randomized trial? Ann Intern Med. 2006;144(11):850–5.
16. **Pauker SG, Kassirer JP.** The threshold approach to clinical decision making. N Engl J Med. 1980;302(20):1109–17.
17. **Santaguida PL, Riley CM, Matchar DB.** Chapter 5: Assessing risk of bias as a domain of quality in medical test studies. J Gen Intern Med. 2012.
18. **Hartmann KE, Matchar DB, Chang S.** Chapter 6: Assessing applicability of medical test studies in systematic reviews. J Gen Internal Med. 2012. doi:10.1007/s11606-011-1961-9.
19. **Trikalinos TA, Kulasingam S, Lawrence WH.** Chapter 10: Deciding whether to complement asystematic review of medical tests with decision making. J Gen Intern Med. 2012.
20. **MacCannell T, Umscheid CA, Agarwal RK, Lee I, Kuntz G, Stevenson KB, the Healthcare Infection Control Practices Advisory Committee.** Guideline for the prevention and control of norovirus gastroenteritis outbreaks in healthcare settings. Infect Control Hosp Epidemiol. 2011;32(10):939–69.
21. **Turcios RM, Widdowson MA, Sulka AC, Mead PS, Glass RI.** Reevaluation of epidemiological criteria for identifying outbreaks of acute gastroenteritis due to norovirus: United States, 1998-2000. Clin Infect Dis. 2006;42(7):964–9.